METHODS IN NEXT-GENERATION SEQUENCING
Review Article • DOI: 10.2478/mngs-2012-0001 • MNGS • 2013 • 1–9

VERSITA

Using NextGen sequencing to identify the molecular basis for genetic disease: an evolutionary perspective

Abstract
In this review, I consider how genetic diseases are related to the evolution of humans and the importance of understanding the results from NextGen sequencing projects in this context. Genome wide association studies (GWAS) were predicated on the assumption that genetic disease was caused by many small effect variations found in large numbers of individuals. However, NextGen sequencing has demonstrated that many disease-causing mutations have a large effect and occur only within restricted populations, families or even individuals. The frequency with which specific types of mutations occur, their effect size and their distribution within the human population is currently an area of active research in genetics. A greater appreciation of human evolutionary history will allow us to design more informative studies to address these questions and properly interpret their results.

Keywords
High-throughput DNA sequencing • Evolution • Mutation • Genetic variation • Genetic diseases • Comparative genomics

Robert B. Norgren, Jr*

Department of Genetics,
Cell Biology and Anatomy,
University of Nebraska Medical Center

Received 03 May 2012
Accepted 23 October 2012

Introduction

NextGen sequencing is used for evolutionary studies [1-3] and to determine the genetic contributions to human disease [4-6]. However, the relationship between evolution and genetic disease is sometimes not considered when interpreting or discussing the results of NextGen sequencing. In this review, I will consider how evolution has shaped human susceptibility to genetic disease and how this is relevant to interpretation of NextGen sequencing results.

Many approaches have been utilized to discover the genetic causes of disease. In the past, family studies used genetic markers in multiple generations of individuals to identify regions in chromosomes which seemed to be inherited by family members who suffered from a disease afflicting some, but not all, members of the family [7]. In the pre-genomics era, this approach required a heroic effort and substantial financial investment due to sequencing costs. A newer approach, termed genome wide association studies (GWAS), promised a less expensive method for quickly identifying large numbers of variations associated with genetic disease [8]. The assumption behind this approach, that many deleterious mutations, each having a small effect but widely distributed in the human population, has been shown to be true in some cases [9]. However, GWAS studies have not been able to detect the genetic cause of disease for most patients despite large studies raising the issue of "missing heritability" [9,10].

Some of the "missing heritability" has been shown to be due to geographical region-specific variation. Although most GWAS studies have drawn from European populations, more recent studies have profitably focused on Africans [11-13] and Asians [14,15]. These studies demonstrate that variations which are associated with disease in a geographical population may not be present in another group or, if they are, may not have the same association with disease.

Many conditions are due to mutations with a low minor allele frequency, i. e., mutations which are present only within restricted groups [6,9,16,17]. Exome sequencing has revealed that the cause of at least some genetic diseases is family or even individual specific [6,18-20]. Specific mutations which are rare and distinctive among individuals and families cannot be found from GWAS studies [7,17]. Hence, it is particularly fortunate that NextGen sequencing is becoming affordable at this time. It will permit comprehensive sequencing, first of exomes, but later of entire genomes of individuals for the cost of a standard medical test.

Framework for discovery

We know that common, small effect variations and rare, large effect variations both contribute to phenotypes [21]. Disease-related variations may be present in four general levels of frequency: 1. across the entire human population, 2. restricted to a geographical region, 3. present only within a particular

* *E-mail: rnorgren@unmc.edu*

family or 4. have just originated in a single individual. At the current time, we do not know what percent of genetic disease is accounted for by variants at each of these four levels of distribution. NextGen sequencing has the potential to identify disease causing mutations and quantify their frequency within the human population.

One of the central problems in applying NextGen sequencing to genetics is determining whether a particular variant present in individuals with a disease *causes* the disease. This difficulty arises because there are so many variations present within each person [22]. Detecting the signal of disease from the noise of random variation requires intelligent filtering. However, determining which type of filter is most appropriate depends greatly on evolutionary context [22].

Genetic variation among individuals can influence disease susceptibility. Since diseases can obviously affect the ability of an individual to survive and reproduce, genetic variations which increase or decrease an individual's susceptibility to disease would be expected to be subject to natural selection.

At a molecular level, there are many different types of possible mutations including: 1. insertions or deletions of one more nucleotides in coding exons or important regulatory regions 2. substitutions of nucleotides in coding sequence which alter amino acids 3. alteration of exon splice sites 4. increase or decrease in the number of copies of a gene. These changes may prevent functional protein from being made, change the amount of protein being made or alter its function.

New genetic mutations can have three possible outcomes: no effect, a disadvantageous effect or an advantageous effect. No effect mutations are termed neutral. They are of no benefit to an individual who carries them, but also have no cost. The frequency of such mutations can vary in different populations due to genetic drift but would not be expected to change rapidly within a given population unless they were in close proximity to a mutation under selection (genetic sweep) [23]. Mutations which are disadvantageous are usually selected against. That is, since individuals with these variations are less likely to have children, the mutation should become less common in a given population over time. This is termed "negative" or "purifying" selection. The more severe the deficit, the more rare should be the observed frequency of disadvantageous mutations. Mutations which are advantageous to the individuals which carry them should become more common over time. This is termed "positive" selection. The more beneficial the mutation, the more rapidly it should spread within the population. Selective pressures can change. For example, mutations will start to accumulate in genes which had been under negative selection but which are no longer important to a species due to a change in the environment the organism is evolving in. An evolutionarily naive interpretation of variations present in such genes may mistakenly attribute high impact to loss-of-function mutations in genes which no longer have much, if any, significance for human health.

The size of a given group, how rapidly it expands in numbers and when and how often interbreeding with other groups occurs will all affect the spread of new mutations within the human population [23,24].

There are three main variables which affect interpretation of NextGen sequencing results with respect to human genetic disease: molecular pathology, influence of selection, and population history. Given the current state of our knowledge, it is too soon to develop a formal theory which can be used in an automated way to ensure correct interpretation of every genetic disease. Many more individuals with well-defined phenotypes and pedigrees need to be sequenced before such a theory can mature. However, in the interim, it is useful to consider specific examples of how the three variables identified here interact to produce a broad array of genetic deficits. The examples which follow were chosen to illustrate the importance of evolutionary thinking in the interpretation of NextGen sequencing results for the diagnosis of genetic disease.

Gain of function disadvantageous variations - trinucleotide repeats

Mutations in one allele of a gene can change its function such that the protein produced becomes harmful. CAG repeats in the Huntingtin gene are of variable length [25,26]. Since CAG codes for glutamine, CAG repeats will code for long chains of gluatamine. In some individuals with many CAG repeats, these long tracts of glutamine create a toxic protein which damages neurons in the caudate nucleus of the brain causing Huntington's disease [25,26]. This disease exhibits genetic anticipation, ie, the repeats may expand from one generation to the next leading to increasingly early and severe disease with each new generation [25]. Since this category of mutation is dominant and disadvantageous, one would expect that it would be under strong negative selection. Huntington's disease is rare. But it may be more common than expected due to the nature of the CAG repeats. The mere presence of this sequence pattern may predispose this gene to new mutations. So, although individuals with severe forms of this gene may be at a selective disadvantage, new CAG repeats in Huntingtin causing this disease in other families may be occurring at a more rapid rate than other types of mutations. Indeed, expanding trinucleotide repeats represents an entire class of genetic lesions involving the following known genes/disorders: ATXN1/spinocerebellar ataxia type 1; ATXN2/spinocerebellar ataxia type 2; ATXN3/Machado-Joseph disease; CACNA1A/spinocerebellar ataxia type 6; ATXN7/spinocerebellar ataxia type 7; ATXN8/spinocerebellar ataxia type 8; AR/spinal bulbar muscular atrophy; PPP2R2B/spinocerebellar ataxia type 12; FRM1/fragile X syndrome; DMPK/myotonic dystrophy [25,26].

CAG repeats prompt two caveats for investigators performing NextGen exome sequencing. First, the issue of genetic anticipation needs to be considered. Parents may or may not show signs of the disease, but one may have "prepathological" changes in a trinucleotide repeat length. Hence, this is a variation which will be expected to change across the generations. Second, mapping, identifying, and

counting the repeat lengths may prove difficult with NextGen short sequences when using alignment-based approaches. Several new sequencing technologies promise longer reads. When these become accurate and cheap, these new methods should greatly ameliorate the problems associated with sequencing repetitive regions. Sequencing entire families should help identify causative mutations. This should improve genetic counseling for these diseases. It may even be possible to identify pre-pathological CAG expansions once costs are low enough for genome sequencing of all individuals who wish it.

Loss of function disadvantageous variations

Mutations in one or both alleles of a gene can cause it to stop functioning. If the affected gene has a useful function, then such mutations would be expected to be under negative selection. In some cases, loss of function in a single allele can cause a deficit (haploinsufficiency) [27]. However, in most cases, loss of function mutations in both alleles are required before a phenotype is observed. Estimates of the percent of new mutations affecting amino acids under negative selection range from 30 - 70% [23]. For variations where the phenotype is severe, one would expect that negative selection should be strong and the incidence of such mutations rare. Two lines of evidence suggest that this is true. First, the incidence of serious disease associated with loss of function mutations *is* rare, approximately 1 in 10,000, for many conditions [28]. Second, independent mutations in the same gene have been observed. This is especially easy to document in large genes like ATM (ataxi-telangiectasia) [29,30] and DMD (muscular dystrophy) [31]. Large genes are overrepresented in genetic diseases because their many exons make them statistically more likely to be subject to mutations which affect protein function.

Ataxia-telangiectasia (AT) is considered an autosomal recessive disease, although there is some evidence that women with one nonfunctioning allele are at increased risk for breast cancer [27]. Children with AT, a severe genetic disorder which causes degeneration of neurons responsible for coordination (Purkinje cells of the cerebellum) and increased risk of cancer, are typically compound heterozygotes with respect to loss of function mutations in the ATM gene [29]. That is, the mutations in their two ATM alleles are different. A large number of different loss of function mutations have been documented among AT patients [29]. This implies that mutiple ATM mutations are arising spontaneously and likely being extinguished as a result of selection rather than that a single mutation is present and being maintained at a constant low level. DMD, the gene for dystrophin, is present on the X chromosome and is also very large [31]. Women who are carriers usually do not exhibit severe symptoms, but their sons experience muscular dystrophy. The severity of this disease depends on the location of the mutation and its effect on dystrophin function. The wide variety of mutations in this gene also argue for many different mutation events which are selected against rather than a single mutation present at a constant incidence within the general population.

Identifying the many different mutations which cause AT and muscular dystrophy has been difficult and expensive. For AT and other rare diseases, proper diagnosis may take years. Although we are in the early stages of cataloging all of the mutations which can cause disease, we already have sufficient evidence from genetic diseases including AT [29], muscular dystrophy [31], Charcot-Marie-Tooth disease [19], Miller syndrome [32] and autism [20,33] to suggest that many families will possess mutations specific to them. Further, some loss of function mutations are new (*de novo*) and hence will not be found in either parent. These mutations can only be identified by sequencing genomic DNA from both an affected child and other family members. NextGen sequencing could be used to identify family and individual specific loss of function mutations rapidly and cheaply. By knowing the specific mutation which causes disease in a family or individual, research targeted at treating the underlying molecular pathology will be possible. There is reason to hope that this will be more effective than treatment designed solely on the basis of presenting symptoms.

Inbreeding increases the odds that offspring will possess two copies of an otherwise rare deleterious loss of function mutation [34]. One would therefore expect an increased incidence of genetic disease in those individuals. In the case of incest and marriage between close relatives, such genetic diseases have been well described. This would be expected to result in fewer surviving children (inbreeding depression). A study of inbred Swiss women supports this hypothesis [35].

The Amish are a religious group settled primarily in North America which originated from a small group of founders (perhaps as little as 200 in Lancaster, PA, the location of the biggest group of Amish) originally from Switzerland [36]. Due to their strict religious practices, almost all marriages occurred among members of the same group. Some genetic diseases have been observed among the Amish at higher rates than the general North American population [37].

Given the proven negative effects of inbreeding on fertility, one might expect that the Amish population has declined from its original 200 members in Lancaster, PA. In fact, in 1960, there were 43,000 members. Some of these may be descendants of additional immigrants, but it is likely that much of this increase was through births in the US. Furthermore, the Amish population has increased dramatically in recent years - 102% from 1991 to 2010 [38]. This recent increase is due to births, not immigration. Amish families often have 5 or more children. Their fertility rate is far above the average for Non-hispanic whites of 1.84 [39,40].

The Amish are well-known for their aversion to medical care [41]. Although they will accept it under some (usually severe) circumstances, they receive considerably less medical care than most Americans. Further, they lead physically hard lives. Finally, they receive a limited education which is normally associated with a shorter life expectancy. Yet, the Amish live about 71 years, on average [42].

There are now almost 250,000 Amish in North America [43,44]. This number is expected to double by 2024 [43,44]. If the current rate of increase of 5% per year is maintained, the Amish

will represent a substantial proportion of the North American population by the end of the century. It is difficult to square the rapid growth in the Amish population with inbreeding depression or reduced fitness. From an evolutionary perspective, the Amish would appear to be spectacularly successful.

Both the high fertility of Amish women and a normal lifespan argue against a population-limiting genetic load in this group. How can this be given the known level of inbreeding? Rebound in fertility after an initial decrease has been noted in many studies of inbred organisms. Getting past this "bottleneck" has been attributed to a purging of deleterious mutations which actually occurs more rapidly in inbred than in outbred populations [45-47]. Thus, although the Amish continue to suffer from genetic disease, many severe loss-of-function mutations may no longer be present within this group. This hypothesis could be tested with NextGen sequencing studies. Is the total genetic load within inbred populations which have survived a bottleneck lower than in outbred populations? Genetic studies in the Amish have demonstrated that the range of deleterious mutations found in a given population vary widely depending on the history of different human groups [37]. NextGen exomic sequencing in this group has already identified loss of function mutations related to genetic disease [37]. The authors of this study describe how their detailed knowledge of the history of this particular group and their familiarity with the different families has helped guide their research project. A similar in-depth approach to genetic medicine in other reproductively isolated groups is likely to yield further insight into the range of disease-causing mutations in different human populations.

Loss of function no effect or advantageous variations

Although counter-intuitive, not all loss of function mutations in genes are deleterious. Many have no effect on fitness despite being graded as "high impact" while other loss of function mutations actually increase fitness.

The olfactory receptor gene family is one of the largest in the mammalian genome - over one thousand genes. However, in humans, many of these genes are no longer functional [48,49]. This pattern of functional olfactory receptor genes becoming pseudogenes has been observed in other primates. It is usually attributed to a relaxation on selection for olfactory function that occurred as primates moved into the trees and began depending more on vision than olfaction [50]. Hence, a change in environment for primates has resulted in a loss of purifying selection on olfactory receptor genes.

In a recent study of NextGen exomic data, approximately 100 high impact loss of function mutations were found in individual humans [6]. These loss of function mutations were not evenly distributed among genes. Loss of function mutations among genes which were highly conserved between rhesus macaques and humans were less likely to be found than mutations in genes with poor conservation between these two species [6]. One interpretation of these results is that highly conserved genes are more likely to be under negative selection than poorly conserved genes. This suggested a potential filter for identifying the loss of function mutations most likely to actually cause disease: conservation of protein sequence between rhesus macaques and humans [6]. Olfactory receptor proteins are poorly conserved between rhesus and humans [49]. Olfactory receptor genes were disproportionately represented on the list of homozygous loss of function mutations [6]. Hence, olfactory receptor genes not yet reported as pseudogenes are likely no longer under purifying selection and may become pseudogenes in the entire human population at some point in the future. Olfactory receptor and other genes which are now under relaxed selection may be mistaken as potential disease related candidates due to the presence of high impact mutations in these genes. By comparing the human genome with high quality nonhuman primate genomes, it should be possible to develop filters to remove genes no longer under selection in humans from genetic disease candidate gene lists.

Because viruses often use host proteins to enter cells, the absence or alteration of these proteins may result in decreased susceptibility to viral disease [51]. The Norwalk virus (Norovirus) is a common cause of gastroenteritis (stomach flu). This virus interacts with the Fut2 protein. However, some individuals do not have functioning FUT2 genes - both alleles have loss of function mutations [52]. People with these mutations are termed "nonsecretors". These individuals do not appear to be susceptible to Norovirus infection. Although there is debate as to whether this gene is under positive selection, a specific inactivating mutation of the FUT2 gene appears to be spreading rapidly within the Asian population which suggests that this mutation confers an advantage on its carriers [52]. There is a small increased risk of diabetes type I associated with nonfunctional FUT2 alleles [53]. However, it may be that the advantage of resistance to norovirus outweighs the disadvantage of increased susceptibility to diabetes type I, perhaps especially in areas with dense populations where nororvirus can easily spread and infect large numbers of people.

The HIV virus causes AIDS. Many strains of HIV interact with the human CCR5 protein to gain entrance into host cells. Loss of function mutations in the CCR5 gene have been shown to confer resistance to HIV pathogenesis in carriers, presumably because this limits the ability of HIV to enter host cells [54,55]. Heterozygotes with one loss of function mutation had a significantly lower risk of AIDS progression after HIV exposure compared to individuals with two functioning alleles. Homozygotes with two loss of function CCR5 alleles appear to be completely protected against AIDS progression after exposure to HIV. Given that HIV has emerged as a viral pathogen in humans only within the last few decades whereas the protective CCR5 allele evolved about 700 years ago [55,56], it seems unlikely that loss of function mutations in CCR5 achieved their current distribution in the human population as a result of selection for HIV protection. Further, no loss of function mutations in CCR5 were found in Africa where HIV originated but were instead found in Europe [57]. Some have argued that this mutation came under

selection to protect against a different virus prevalent in Europe [56] (see [58] for an opposing view). The fact that this mutation also protects against HIV may indicate a similar viral strategy for host cell invasion.

The two examples above illustrate two different evolutionary principles. First, the fact that different FUT2 mutations protective against norovirus are found in different populations suggest convergent evolution in response to similar selection pressure. The fact that a mutation protective against HIV already existed in the European population before HIV arose as a threat suggests there may be a limited number of host proteins which viruses can use to infect humans. Thus, some humans may be protected against pathogens which have not yet emerged. As we do not yet know the full range of advantageous loss of function mutations, caution should be exercised when interpreting "high impact" loss-of-function mutations, especially in reference to genes which may interact with infectious agents.

Loss of function heterozygote advantage/ homozygote disadvantage variations

Some variations in genes can confer an advantage to an individual if only one allele is affected but a disadvantage if two alleles are present. Two diseases involving hemoglobin illustrate this phenomenon: Sickle cell anemia [59] and the beta thalassemias [60]. In both cases, loss of function mutations in a single allele protect carriers against malaria. However, loss of function mutations in both alleles can result in severe disease and reduced life expectancy. The mutation associated with sickle cell anemia occurs mainly in Africa while the beta thalessemia mutations occur in Europe and Asia - another example of convergent evolution. As a result of the benefit to heterozygotes, these alleles and the genetic diseases they are associated with are found at higher levels than would otherwise be predicted in areas where malaria is endemic.

Emergent disadvantageous and advantageous variations

Some variations associated with human genetic diseases actually represent the ancestral or wild type state. Thus, the variations associated with disease for this class of variations are not due to mutations. For these diseases, only humans with new variants, not present in other species, will have decreased susceptibility to what otherwise would likely be very common genetic diseases. The explanation for this counter-intuitive suggestion is that humans have recently come under new selective pressures due to changes in physiology, environment or behavior. Variations which were previously neutral or perhaps even advantageous have suddenly become disadvantageous.

There are multiple genetic and environmental factors which can influence the odds of getting Alzheimer's disease, a condition which causes portions of the cerebral cortex of the brain to degenerate and results in dementia. The odds of getting late onset dementia has been shown to be strongly influenced by which apolipoprotein E (APOE) alleles an individual has [61]. There are three alleles: 2, 3 and 4. The odds of getting late onset Alzheimer's are estimated to be: APOE2/2: 0.08%; APOE2/3 - 3.2%; APOE3/3 - 5.1%; APOE3/4 - 18%; APOE4/4 - 67%. Without an evolutionary context, it would be tempting to consider APOE4 (and possibly APOE3) as disease causing mutations. In fact, APOE4 is the ancestral allele possessed by all mammals [62]. It is APOE3 and APOE2 which are the mutations, apparently unique to humans. The explanation appears to be that since the negative effects of APOE4 on cognition are primarily apparent after age 50 [63,64], for shorter-lived ancestral species, APOE4 had no disadvantageous effects. APOE2 and APOE3 appear to be new alleles under positive selection in humans. One possible explanation for the apparent selection in post-reproductive individuals is that family groups with cognitively intact older individuals will be able to draw on that experience to increase the odds that related children will survive infancy - the Grandmother hypothesis [65-67].

Craig Venter, one of the sources for the human genome project, is APOE3/4 [68]. Thus, a NextGen sequence which agrees with the reference genome, and which is conserved across species, may nonetheless be associated with disease in humans. It is likely that other "reference" sequences may in fact be associated with disease. Care should be taken when describing whether a variation is "normal" or "mutated". The possibility that "mutation" may be associated with the absence of disease rather than its presence should be considered.

The ability to drink milk as an adult is determined by variations in the lactase (LCT) gene. Although treated as a disease, the allele associated with lactose intolerance is ancestral [69]. The ability to drink milk as an adult is the result of a mutation which has spread primarily in the northern European population within the last 20,000 years [69]. It appears to be under positive selection in this population as a result of a dependence on dairy products. The rise of dairying created a new selection pressure. Possessing the LCT allele, which resulted in persistence of lactase in adults, conferred an advantage to individuals with access to dairy products. In an example of convergent evolution, a different mutation which results in the persistence of lactase appears to be under positive selection in a group of East African who raise dairy animals [70]. This example suggests that defining disease requires more than a comparison of sequences. Individuals with the ancestral form of LCT will not experience disease if they do not consume dairy products as adults.

There is a growing list of diseases associated with exposure to environments different from those experienced by an individual's ancestors. Individuals with relatively little pigmentation are at greatly increased risk of skin cancer if they choose to live near the equator [71]. Individuals who are heavily pigmented are at greatly increased risk of vitamin D deficiency if they live at high latitudes [72]. Among humans who live where their ancestors evolved, pigmentation is highly correlated with latitude [71]. This is because natural selection for skin cancer versus vitamin D deficiency act in opposite directions and are relatively stronger at different latitudes [71,73]. There are likely other variations which

are adaptive in one environment but deleterious in another. Thus, consideration of the particular evolutionary history and current environment of individual patients may be key to understanding whether a particular sequence variation is associated with disease.

The genomic DNA of multiple individuals from different geographic origins (Africans, Asians and Europeans) was mixed together before sequencing the human reference genome [74,75]. Thus, when one is aligning NextGen sequences against the human reference genome, one may be aligning against an individual from any of several different geographic origins. This may complicate interpretations of how significant a variation is. A difference between the reference sequence and a patient sample from the same geographical region may suggest a causative mutation while the same variation between the reference sequence and a patient sample from a different geographical region may have no clinical significance. To determine whether this potential problem is actually confounding interpretation, it would be necessary to label the different parts of the human reference genome by individual contributor or at least by geographical origin. Alternatively, high quality human reference genomes from separate individuals representing multiple geographic regions could be produced. The individual genomes produced thus far fall far short of what would be necessary for a high quality reference genome.

Since a disproportionate number of studies of genetic variations have focused on individuals with ancestors who evolved in Europe, there is relatively less known about which genetic variations contribute to disease in other groups. NextGen sequencing could be used to rapidly increase the catalog of such mutations throughout the human population. The 1,000 genomes project is an early step in this direction [76].

Nonhuman primate (NHP) genomes, human evolution and genetic disease

Some of the analyses related to natural selection in humans rely on data from the chimpanzee and rhesus monkey genomes. However, we have determined that draft mammalian genomes such as the rhesus have sequencing and annotation errors which affect 50% of genes [77]. Thus, the true level of natural selection acting on coding exons and regulatory regions is difficult to determine with precision at this time. Efforts are under way to improve the quality of draft NHP genomes. Better NHP genomes will also provide more accurate filters when determining whether a “high impact” mutation is actually occurring in an important gene or whether it has little relevance because the gene is no longer under selection in humans.

Summary

Human genetic disease can be properly understood only in the context of natural selection. In some ways, biology is more similar to history than physics or chemistry. The latter disciplines are amenable to the assumption that the rules that apply in one situation will be generalizable to other situations. This is not the case for biology - especially genetics. Every mutation has its own story. Some mutations are deeply rooted in human evolution and define us as a species. Others have just occurred *de novo* in a single individual. Most mutations came into existence somewhere between these two extremes. Are they spreading within the human population or decreasing in frequency? This is a question evolutionary biologists ask but is highly relevant to clinical genetics as well. NextGen sequencing can answer this question because we can, in theory, determine the history of every mutation by sequencing large numbers of human genomes. Falling costs mean that such a project is not only feasible, but likely to happen in the near future. But interpretation of the results is critically dependent on an appreciation of the evolutionary forces that shaped our species in the past and influence it to this day.

Outlook

NextGen sequencing has opened up a remarkable opportunity to discover the variations in the human genome associated with genetic disease. This technology can be usefully informed by evolutionary biology when evaluating potentially deleterious variations in humans. However, it may also help uncover evidence of selection in humans which may ultimately provide candidate genes for therapeutic targets.

References

[1] Elmer K.R., Meyer A., Adaptation in the age of ecological genomics: insights from parallelism and convergence, Trends Ecol. Evol., 2011, 26, 298-306

[2] Stapley J., Reger J., Feulner P.G., Smadja C., Galindo J., Ekblom R., Bennison C., Ball A.D., Beckerman A.P., Slate J., Adaptation genomics: the next generation, Trends Ecol Evol., 2010, 25, 705-712

[3] Stoneking M., Krause J., Learning about human population history from ancient and modern genomes, Nat. Rev. Genet., 2011, 12, 603-614

[4] Nelen M., Veltman J.A., Genome and exome sequencing in the clinic: unbiased genomic approaches with a high diagnostic yield, Pharmacogenomics, 2012, 13, 511-514

[5] Gonzaga-Jauregui C., Lupski J.R., Gibbs R.A., Human genome sequencing in health and disease, Annu. Rev. Med., 2012, 63, 35-61

[6] MacArthur D.G., Balasubramanian S., Frankish A., et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science, 2012, 335, 823-828

[7] Bailey-Wilson J.E., Wilson A.F. Linkage analysis in the next-generation sequencing era, Hum Hered. 2011, 72, 228-236

[8] Lander, E.S., The New Genomics: Global views of biology, Science, 1996, 274, 536-539

[9] Manolio T.A., Collins F.S., Cox N.J., Goldstein D.B., Hindorff L.A., Hunter D.J., McCarthy M.I., Ramos E.M., Cardon L.R., Chakravarti A., Cho J.H., Guttmacher A.E., Kong A., Kruglyak L., Mardis E., Rotimi C.N., Slatkin M., Valle D., Whittemore A.S., Boehnke M, Clark A.G., Eichler E.E., Gibson G., Haines J.L., Mackay T.F., McCarroll S.A., Visscher P.M. Finding the missing heritability of complex diseases, Nature, 2009, 461, 747-753

[10] Maher B. Personal genomes: The case of the missing heritability, Nature, 2008, 456, 18-21

[11] Bustamante C.D., Burchard E.G., De la Vega F.M., Genomics for the world, Nature, 2011,75,163-165

[12] Chen G., Ramos E., Adeyemo A., Shriner D., Zhou J., Doumatey A.P., Huang H., Erdos M.R., Gerry N.P., Herbert A., Bentley A.R., Xu H., Charles BA., Christman M.F., Rotimi C.N., UGT1A1 is a major locus influencing bilirubin levels in African Americans, Eur. J. Hum. Genet., 2012, 20, 463-468

[13] Cabral W.A., Barnes A.M., Adeyemo A., Cushing K., Chitayat D., Porter F.D., Panny S.R., Gulamali-Majid F., Tishkoff S.A., Rebbeck T.R., Gueye S.M., Bailey-Wilson J.E., Brody L.C., Rotimi C.N., Marini J.C., A founder mutation in LEPRE1 carried by 1.5% of West Africans and 0.4% of African Americans causes lethal recessive osteogenesis imperfecta, 2012, Genet Med.

[14] Fu J., Festen E. A., Wijmenga C., Multi-ethnic studies in complex traits, Hum. Mol. Genet., 2011 20, R206-R213

[15] Sim X., Ong R.T., Suo C., Tay W.T., Liu J., Ng D.P., Boehnke M., Chia K.S., Wong T.Y., Seielstad M., et al., Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia, PLoS Genet., 2011, 7, e1001363

[16] McClellan J., King M.C., Genetic heterogeneity in human disease, Cell, 2010, 141, 210-217

[17] Gravel S., Henn B.M., Gutenkunst R.N., Indap A.R., Marth G.T., Clark A.G., Yu F., Gibbs R.A., 1000 Genomes Project, Bustamante C.D., Demographic history and rare allele sharing among human populations, Proc. Natl. Acad. Sci. USA, 2011, 108, 11983-11988

[18] Bamshad M.J., Ng S.B., Bigham A.W., Tabor H.K., Emond M.J., Nickerson D.A., Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery, Nat. Rev. Genet., 2011, 12, 745-55

[19] Lupski J. R., Reid J. G., Gonzaga-Jauregui C., Rio Deiros D., Chen D. C., Nazareth L., Bainbridge M., Dinh H., Jing C., Wheeler D. A., McGuire A. L., Zhang F., Stankiewicz P., Halperin J. J., Yang C., Gehman C., Guo D., Irikat R. K., Tom W., Fantin N. J., Muzny D. M., Gibbs R. A., Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy, N. Engl. J. Med., 2010, 362, 1181-1191

[20] Muers M., Human genetics: Fruits of exome sequencing for autism, Nat. Rev. Genet., 2012, 13, 377

[21] Gibson G., Rare and common variants: twenty arguments., Nat Rev Genet. 2012, 13, 135-145

[22] Durbin R. M., Altshuler D., Abecasis G. R., Bentley D. R., Chakravarti A., Clark A. G., Collins F. S. et al., A map of human genome variation from population-scale sequencing, Nature, 2010, 467, 1061- 1073

[23] Nielsen R., Hellmann I., Hubisz M., Bustamante C., Clark A. G., Recent and ongoing selection in the human genome, Nat. Rev. Genet., 2007, 8, 857-868

[24] Goldstein D. B., Chikhi L., Human migrations and population structure: what we know and why it matters, Ann. Rev. Genomics Hum. Genet., 2002, 3, 129-152

[25] Orr H.T., Zoghbi H.Y., Trinucleotide repeat disorders, Annu. Rev. Neurosci., 2007, 30, 575-621

[26] Shao J., Diamond M,I., Polyglutamine diseases: emerging concepts in pathogenesis and therapy. Hum. Mol. Genet., 2007,2:R115-R123

[27] Thompson D., Duedal S., Kirner J., McGuffog L., Last J., Reiman A., Byrd P., Taylor M., Easton D.F., Cancer risks and mortality in heterozygous ATM mutation carriers. J Natl Cancer Inst. 2005, 97, 813-822

[28] Willems P.J., Bottlenecks in molecular testing for rare genetic diseases, Hum. Mutat., 2008, 29, 772-775

[29] Concannon P., Gatti R.A., Diversity of ATM gene mutations detected in patients with ataxia-telangiectasia, Hum. Mutat., 1997, 10, 100-107

[30] Savitsky K., Bar-Shira A., Gilad S., Rotman G., Ziv Y., Vanagaite L., Tagle D. A., Smith S., Uziel T., Sfez S., Ashkenazi M., Pecker I., Frydman M., Harnik R., Patanjali S. R., Simmons A., Clines G.A., Sartiel A., Gatti R.A., Chessa L., Sanal O., Lavin M. F., Jaspers N. G., Taylor A. M., Arlett C. F., Miki T., Weissman S. M., Lovett M., Collins F. S., Shiloh Y., A single ataxia telangiectasia gene with a product similar to PI-3 kinase, Science, 1995, 268, 1749-1753

[31] Soltanzadeh P., Friez M.J., Dunn D., von Niederhausern A., Gurvich O.L., Swoboda K.J., Sampson J.B., Pestronk A., Connolly A.M., Florence J.M., Finkel R.S., Bönnemann C.G., Medne L., Mendell J.R., Mathews K.D., Wong B.L., Sussman M.D., Zonana J., Kovak K., Gospe S.M. Jr., Gappmaier E., Taylor L.E., Howard M.T., Weiss R.B., Flanigan K.M., Clinical and genetic characterization of manifesting carriers of DMD mutations, Neuromuscul. Disord. 2010, 20, 499-504

[32] Ng S. B., Buckingham K. J., Lee C., Bigham A. W., Tabor H. K., Dent K. M., Huff C. D., Shannon P. T., Jabs E. W., Nickerson D. A., Shendure J., Bamshad M. J., Exome sequencing identifies the cause of a mendelian disorder, Nat. Genet., 2010, 42, 30-35

[33] Miles J. H. Autism spectrum disorders--a genetics review, Genet. Med. 2011, 13, 278-294

[34] Charlesworth D., Willis J. H., The genetics of inbreeding depression, Nat. Rev. Genet., 2009, 10, 783-796

[35] Postma E, Martini L, Martini P., Inbred women in a small and isolated Swiss village have fewer children, J. Evol. Biol., 2010, 23, 1468-1474

[36] McKusick V. A., Hostetler J. A., Egeland J. A., Genetic studies of the Amish. Background and potentialities, Bull. Johns Hopkins Hosp., 1964, 115, 203-222

[37] Puffenberger E. G., Jinks R. N., Sougnez C., Cibulskis K., Willert R. A., Achilly N. P., Cassidy R. P., Fiorentini C. J., Heiken K. F., Lawrence J. J., Mahoney M. H., Miller C. J., Nair D. T., Politi K. A., Worcester K. N., Setton R. A., Dipiazza R., Sherman E. A., Eastman J. T., Francklyn C., Robey-Bond S., Rider N. L., Gabriel S., Morton D. H., Strauss K. A., Genetic mapping and exome sequencing identify variants associated with five novel diseases, PLoS One, 2012, 7, e28936

[38] Young Center for Anabaptist and Pietist Studies, Elizabethtown College. "Amish Population Trends 1991-2010, 20-Year Highlights." http://www2.etown.edu/amishstudies/Population_Trends_1991_2010.asp

[39] Miller K., Yost B., Flaherty S., Hillemeier M. M., Chase G. A., Weisman C. S., Dyer A. M., Health status, health conditions, and health behaviors among Amish women. Results from the Central Pennsylvania Women's Health Study (CePAWHS), Womens Health Issues, 2007, 17, 162-171

[40] Hamilton, B. E., Martin, J. A., Ventura, S. J. Births: Preliminary Data for 2007, National Vital Statistics Reports 57:12, March 18, 2009

[41] Adams C. E., Leverland M. B., The effects of religious beliefs on the health care practices of the Amish, Nurse Pract. 1986, 11, 58, 63, 67

[42] Mitchell B. D., Hsueh W. C., King T. M., Pollin T. I., Sorkin J., Agarwala R., Schäffer A. A., Shuldiner A. R., Heritability of life span in the Old Order Amish, Am. J. Med. Genet., 2001, 102, 346-352

[43] Berg N. Why the Amish Population Is Exploding, The Atlantic Cities, Aug 01, 2012

[44] Golgowski, N., Amish population booming in the U.S. with a new settlement founded nearly once a month, Daily Mail, July 28, 2012

[45] Crnokrak P., Barrett S. C. Perspective: purging the genetic load: a review of the experimental evidence. Evolution, 2002, 56, 2347-2358

[46] Glémin S. How are deleterious mutations purged? Drift versus nonrandom mating, Evolution, 2003, 57, 2678-2687

[47] Reed F. A., Aquadro, C. F. Mutation, selection and the future of human evolution, Trends Genet., 2006, 22, 479-484

[48] Kambere M.B., Lane R.P. Co-regulation of a large and rapidly evolving repertoire of odorant receptor genes, BMC Neurosci. 2007, 8 Suppl 3, S2

[49] Gilad Y., Man O., Pääbo S., Lancet D., Human specific loss of olfactory receptor genes, Proc Natl Acad Sci U S A. 2003,100, 3324–3327

[50] Niimura Y., Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents, Hum. Genomics, 2009, 4, 107-118

[51] Olson M.V. When less is more: gene loss as an engine of evolutionary change, Am. J. Hum. Genet. 1999, 64, 18-23

[52] Ferrer-Admetlla A., Sikora M., Laayouni H., Esteve A., Roubinet F., Blancher A., Calafell F., Bertranpetit J., Casals F., A natural history of FUT2 polymorphism in humans, Mol. Biol. Evol., 2009, 26, 1993-2003

[53] Smyth D.J., Cooper J.D., Howson J.M., Clarke P., Downes K., Mistry T., Stevens H., Walker N.M., Todd J.A., FUT2 nonsecretor status links type 1 diabetes susceptibility and resistance to infection, Diabetes, 2011, 60, 3081-3084

[54] Zimmerman P.A., Buckler-White A., Alkhatib G., Spalding T., Kubofcik J., Combadiere C., Weissman D., Cohen O., Rubbert A., Lam G, Vaccarezza M., Kennedy P.E, Kumaraswami V., Giorgi J.V., Detels R., Hunter J., Chopek M., Berger E.A., Fauci A.S., Nutman T.B., Murphy P.M. (1997) Inherited resistance to HIV-1 conferred by an inactivating mutation in CC chemokine receptor 5: studies in populations with contrasting clinical phenotypes, defined racial background, and quantified risk, Mol Med., 1997, 3, 23-36

[55] Novembre J., Galvani A. P., Slatkin M., The geographic spread of the CCR5 Delta32 HIV-resistance allele. PLoS Biol., 2005, 3, e339

[56] Stephens J.C., Reich D.E., Goldstein D.B., Shin H.D., Smith M.W., Carrington M., Winkler C., Huttley G.A., Allikmets R., Schriml L., Gerrard B., Malasky M., Ramos M.D., Morlot S., Tzetis M., Oddoux C., di Giovine F.S., Nasioulas G., Chandler D., Aseev M., Hanson M., Kalaydjieva L., Glavac D., Gasparini P., Kanavakis E., Claustres M., Kambouris M., Ostrer H., Duff G., Baranov V., Sibul H., Metspalu A., Goldman D., Martin N., Duffy D., Schmidtke J., Estivill X., O'Brien S.J., Dean M., Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes, Am. J. Hum. Genet. 1998, 62,1507-1515

[57] Libert F., Cochaux P., Beckman G., Samson M., Aksenova M., Cao A., Czeizel A., et al, The Dccr5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in northeastern Europe, Hum. Mol. Genet., 1998, 7, 399–406

[58] Sabeti P. C., Walsh E., Schaffner S. F., Varilly P., Fry B., Hutcheson H. B., Cullen M., Mikkelsen T. S., Roy J., Patterson N., Cooper R., Reich D., Altshuler D., O'Brien S., Lander E. S., The case for selection at CCR5-Delta32, PLoS Biol., 2005, 3, e378

[59] Aidoo M., Terlouw D.J., Kolczak M.S., McElroy P.D., ter Kuile F.O., Kariuki S., Nahlen B.L., Lal A.A., Udhayakumar V., Protective effects of the sickle cell gene against malaria morbidity and mortality, Lancet, 2002, 359, 1311-1312

[60] Pattanapanyasat K., Yongvanitchit K., Tongtawe P., Tachavanich K., Wanachiwanawin W., Fucharoen S., Walsh D.S., Impairment of Plasmodium falciparum growth in thalassemic red blood cells: further evidence by using biotin labeling and flow cytometry, Blood, 1999, 93, 3116-3119

[61] Raber J., Huang Y., Ashford J. W., ApoE genotype accounts for the vast majority of AD risk and AD pathology, Neurobiol. Aging, 2004, 25, 641-650

[62] Hanlon C.S., Rubinsztein D.C., Arginine residues at codons 112 and 158 in the apolipoprotein E gene correspond to the ancestral state in humans, Atherosclerosis 1995, 112, 85–90

[63] Caselli R.J., Age-related memory decline and apolipoprotein E e4, Discov. Med., 2009, 8, 47-50

[64] Caselli R.J., Dueck A.C., Osborne D., Sabbagh M.N., Connor D.J., Ahern G.L., Baxter L.C., Rapcsak S.Z., Shi J., Woodruff B.K., Locke D.E., Snyder C.H., Alexander G.E., Rademakers R, Reiman E.M., Longitudinal modeling of age-related memory decline and the APOE epsilon4 effect. N. Engl. J. Med., 2009, 361, 255-263

[65] Williams G.C., Pleiotropy, natural selection, and the evolution of senescence, Evolution, 1957, 11, 398-411

[66] Hamilton W. D., The moulding of senescence by natural selection, J. Theoret. Biol., 1966, 12, 12-45.

[67] Blurton Jones N.G., Hawkes K., O'Connell J.F., Antiquity of postreproductive life: are there modern impacts on hunter-gatherer postreproductive life spans?, Am J Hum Biol., 2002, 14, 184-205

[68] Nyholt D. R., Yu C-E., Vlsscher P. M., On Jim Watson's APOE status: genetic information is hard to hide. European J. Human Genet., 2009, 17, 147-150.

[69] Burger J., Kirchner M., Bramanti B., Haak, W., Thomas, M.G., Absence of the lactase-persistence-associated allele in early Neolithic Europeans, Proc. Natl. Acad. Sci. USA, 2007, 104, 3736-3741

[70] Tishkoff S. A., Reed F. A., Ranciaro A., Voight B. F., Babbitt C. C., Silverman J. S., Powell K., Mortensen H. M., Hirbo J. B., Osman M., Ibrahim M., Omar S. A., Lema G., Nyambo T. B., Ghori J., Bumpstead S., Pritchard J. K., Wray G. A., Deloukas P., Convergent adaptation of human lactase persistence in Africa and Europe, Nature Genet. 2007, 39, 31-40

[71] Parra E.J., Human pigmentation variation: evolution, genetic basis, and implications for public health, Am. J. Phys. Anthropol. 2007, Suppl 45, 85-105

[72] Harris S.S., Vitamin D and African Americans, J. Nutr., 2006, 136, 1126-1129

[73] Jablonski, N.G., The evolution of human skin colouration and its relevance to health in the modern world, J. R. Coll. Physicians Edinb. 2012, 42, 58-63

[74] Lander E. S., Linton L. M., Birren B., et al, International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. Nature, 2001, 409, 860-921

[75] Venter, J. C., Adams, M. D., Myers, E. W., et al, The sequence of the human genome, Science, 2001, 291, 1304-1351

[76] 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, Nature, 2010, 467, 1061-1073

[77] Zhang X., Goodsell J., Norgren R. B. Jr. Limitations of the rhesus macaque draft genome assembly and annotation, BMC Genomics, 2012, 13, 206.